1 Supplementary post hoc statistical analysis

1.1 Evaluation units and hypothesis inventory

Supplementary Friedman and two-sided Wilcoxon signed-rank tests were introduced following the original TEST evaluation to further examine whether the performance differences observed in participant-level outcomes were congruent [1, 2]. Whilst the preceding accuracy and bootstrap analyses characterized predictive performance and its associated uncertainty, these supplementary tests provide a complementary assessment of the evidence for differences between representations. Importantly, these are post hoc analyses of the previously saved evaluation outcomes, conditional on the fitted classifiers, fized developmental decisions, and selected participant cohorts. No models were retrained, nor were the TEST outcomes used to revisit representation selection or hyperparameter tuning.

Similar to the paper, we first establish the statistical unit of comparison, let $u_{m,r,c,p}$ denote the proportion of correctly classified images belonging to participant p, for classifier m, representation r, and configuration c. Consistent with the main evaluation protocol, detection failures, target-matching failures, feature-construction failures, and prediction failures are retained in the accuracy denominator and counted as incorrect classifications. Consequently, the resulting participant-level scores reflect the end-to-end classification pipeline rather than performance conditional on successful landmark extraction. For clean TEST, these scores are computed over all $N = 1 = 000$ participants and their $15,298$ selected images. The perturbation analyses instead use the fixed robustness TEST subset, comprising $N = 500$ participants and $2,242$ selected images. An additional consideration arises for the corruption experiments, where each perturbation family contains two severity settings and three prescribed realizations per setting. Treating these outcomes as independent statistical observations would artificially increase the sample size, since the same participant and selected images appear repeatedly under different perturbation conditions. To avoid this, we first aggregate the outcomes within each participant. Specifically, for perturbation family f, we define

$$u_{m,r,f,p} = \frac{1}{6} \sum_{s=1}^{2} \sum_{d=1}^{3} a_{m,r,(f,s,d),p}, \tag{1}$$

where $a_{m,r,(f,s,d),p}$ denotes the participant-level accuracy under severity s and prescribed realization d. The averaging assigns equal weight to the six evaluated conditions within each family. For Gaussian blur, exposure changes, and JPEG compression, the perturbations are deterministic at a fixed severity, such that the three realizations produce identical outcomes. In contrast, the three occlusion realizations correspond to different prescribed placements of the occluding rectangle. In either case, the settings and realizations are not treated as independent replications. Each participant therefore contributes a single aggregated outcome per classifier, representation, and corruption family, preserving the participant as the fundamental statistical unit. Since all nine representations are evaluated on the same participants and selected images within each context, the resulting observations are naturally paired. We exploit this structure through a two-stage description of the comparisons, beginning with an omnibus assessment across the complete representation pool and ending with an individual representation comparison. For each classifier, a Friedman test compares the nine representations on clean TEST and separately within each of the four corruption families, resulting in $2 \times 5 = 10$ omnibus hypotheses. These tests investigate whether the representation rankings differ systematically across participants, without presupposing which particular representations might account for any observed differences. However, an omnibus rejection does not identify the representations responsible for the discrepancy. We therefore consider all $\binom{9}{2} = 36$ representation pairs relative to each classifier, yielding $2 \times 5 \times 36 = 360$ two-sided Wilcoxon signed-rank hypotheses. Four additional clean TEST comparisons examine the removal of normalized spatial coordinates or normalized distances from the full hybrid, separately for each classifier. The corresponding angle-removal comparison is

already represented by the full versus simple hybrid evaluation and is consequently not counted a second time. Thus, the supplementary analysis comprises of 364 paired hypotheses and 10 omnibus hypotheses, for a total of 374 tests.

1.2 Rank statistics, ties, and multiple comparisons

For the omnibus comparisons, the Friedman test operates on the relative performance of the nine representations on eacg participant, rather than directly comparing their pooled classification accuracies. Specifically, the nine participant-level scores are ranked, assigning rank one to the representation attaining the highest accuracy and average ranks whenever two or more representations obtain identical outcomes. This procedure accounts for the paired evaluation structure, since every participant contributes a corresponding score for each representation. A tie-corrected Friedman statistic Q is then evaluated using the chi-square approximation with $k-1=8$ degrees of freedom, where $k=9$ denotes the number of representations [1]. The tie correction is curcial since several geometric representations can produce near identical participant-level accuracies.

Alongside the omnibus significance probabilities, we report Kendall's coefficient as a measure of the consistency of the representation rankings across participants. For the tie-corrected statistic, this is computed as

$$W = \frac{Q}{N(k-1)} = \frac{Q}{8N}, \tag{2}$$

where N denotes the number of participants in the corresponding evaluation context. Whilst the Friedman statistic provides evidence concerning differences amongst the representations, Kendall's W additionally characterizes the degree to which participants exhibit consistent relative rankings. The two quantities therefore provide complementary descriptions of the omnibus comparison, although neither identifies which individual representation pairs differ. To investigate those individual differences, we employ two-sided Wilcoxon signed-rank tests on the paired participant-level scores [2]. For representations r and r', the paired difference is defined as

$$d_p = u_{m,r,c,p} - u_{m,r',c,p},$$

where a positive value indicates higher participant-level accuracy for r, and a negative value favors r'. Unlike a comparison of unpaired accuracy distributions, this construction preserves the fact that the two representations were evaluated on the same participant. The Wilcoxon procedure then ranks the magnitudes of the nonzero paired differences whilst retaining their signs, thereby assessing evidence for a systematic signed difference without restricting to a particular direction. Now since the scores could differ only through numerical precision, the inputs to the rank calculations are rounded to twelve decimal places to suppress numerically consistent ties. This operation affects only the construction of the rank statistics; the underlying reported accuracy estimates and bootstrap confidence intervals remain unchanged. Participants with exactly zero paired differences are omitted from the signed-rank statistic because they favor neither representation, whereas nonzero differences with identical absolute magnitudes receive average ranks, or midranks. Importantly, omitting zero differences from this statistic does not remove the corresponding participants or images from the original classification accuracy calculations. The calculation of the two-sided significance probabilities additionally depends on the number of nonzero paired differences, denoted by n_{nz}. When $1 \leq n_{\mathrm{nz}} \leq 50$ we evaluate the exact sign distribution oc the observed midranks. For larger comparisons, where exact enumeration becomes computationally more demanding, we instead employ a normal approximation incorporating corrections for tied ranks and continuity. Comparisons containing no nonzero paired differences are assigned $p=1$, since the observed outcomes provide no evidence against the null hypothesis. This treatment ensures that ties and all-zero comparisons are handled explicitly rather than being inadvertently interpreted as evidence of a difference. Finally, because the

analysis considers numerus representation comparison, interpreting each unadjusted significance probability independently would increase the risk of false rejections. We therefore apply a single Holm step-down correction across all 374 hypotheses, comprising the ten Friedman omnibus tests and the 364 distinct Wilcoxon comparisons [3]. Statistical significance is assessed at the family-wise level of 0.05, with a hypothesis rejected only when its Holm-adjusted probability satisfies $p_{\text{Holm}} \leq 0.05$. All pairwise comparisons are reported regardless of whether their corresponding omnibus test rejects, ensuring that the complete comparison inventory remains available for inspection. Extremely small chi-square tail probabilities are evaluated in the log domain to avoid numerical underflow and are reported as bounds rather than being represented as zero.
It is important to distinguish these supplementary hypothesis tests from the bootstrap analysis employed in the main manuscript. The latter constructs individual, unadjusted 95% percentile confidence intervals using $10,000$ participant-block bootstrap resamples and therefore quantifys uncertainty in participant-balanced accuracy and mean paired performance differences. The Wilcoxon tests, however, examines the signed-rank distribution of participant-level differences, whilst the Friedman tests assess their relative rankings across the representation pool. Consequently, their conclusions need not coincide, particularly after correcting the supplementary hypotheses for multiple comparisons.

1.3 Results

The Friedman tests reject in all ten classifier combinations after the common Holm adjustment (Table 1). The observed Kendall's W ranges from 0.112 to 0.303, indicating nonidentical participant relative representation ranks, without implying that any one representation is superior. Of the 364 distinct Wilcoxon tests, 289 reject after Holm adjustment: 285 of the 360 representation pairs and all four additional clean ablation comparisons (ϕ_f versus $\phi_{f\setminus A}$ and $\phi_{f\setminus\delta}$ for each classifier). Amongst the 364 tests, 34 use the exact conditional midrank distribution and 330 use the normal approximation. The complete 360 pooled-representation adjusted probabilities and mean participant ranks are reported in the matrices below; the four additional clean ablation comparisons are given separately in Table 3.
In the main paper we discuss that the full and simple hybrid comparisons do not reject on clean TEST or in any corruption family Table 2),in particular, the random-forest comparison yields a positive 0.231-percentage point clean difference and a positive 0.223 percentage point Gaussian-blur difference in the original estimates, with unadjusted bootstrap intervals excluding zero. Nevertheless, their globally adjusted Wilcoxon probabilities are 0.622 and 1.000, respectively. This is not a contradiction: the bootstrap intervals address mean paired accuracy differences without multiplicity adjustment, whereas the post hoc signed-rank tests address the distribution of signed participant differences under a different null hypothesis.
The signed-rank test's null hypothesis concerns the sign distribution of paired participant-level differences rather than the mean paired difference quantified by the bootstrap. Many participant pairs have zero differences, so the number of nonzero pairs can be much smaller than the nominal cohort size; both are reported for the full and simple comparisons. The same participants recur across representation comparisons and corruption families, and the different contexts are not independent datasets. The reported inferences remain conditional on the fixed fitted models, selected participants, and analysis choices, without incorporating uncertainty from model retraining, hyperparameter selection, or representation selection. Crucially, failure to reject a pairwise comparison does not establish statistical or geometric equivalence.

1.4 Complete paired comparisons among the nine representations

The following ten matrices report all 360 Wilcoxon comparisons among the nine pooled representations. Their upper triangles contain the Holm-adjusted probabilities and their second columns give the average participant ranks, where rank one is best. A dash denotes a duplicate

or self-comparison; values below 10^{-4} are displayed as bounds, not as zero. The representations follow the notation and order of the main manuscript: ϕ_q (raw coordinates), ϕ_T (translated coordinates), ϕ_A (axis-normalized coordinates), ϕ_d (distances), ϕ_δ (normalized distances), ϕ_θ (angles), ϕ_s (simple hybrid), ϕ_f (full hybrid), and ϕ_Z (centered isotropically normalized coordinates). Each off-diagonal upper-triangle entry represents one distinct hypothesis. The four additional clean ablation tests appear in Table 3.

Table 1: Tie-corrected Friedman tests for nine representations in each classifier–context combination. Q uses eight degrees of freedom, and $W = Q/(8N)$. Holm-adjusted probabilities are from the common family of 374 tests; values smaller than machine precision are given as bounds.

Classifier	Context	N	Q	Kendall's W	Holm p
LR	Clean	1,000	1746.339	0.218	$< 10^{-300}$
LR	Blur	500	728.349	0.182	2.08×10^{-149}
LR	Exposure	500	484.761	0.121	4.78×10^{-97}
LR	JPEG	500	466.442	0.117	4.01×10^{-93}
LR	Occlusion	500	448.512	0.112	2.77×10^{-89}
RF	Clean	1,000	2423.467	0.303	$< 10^{-300}$
RF	Blur	500	770.520	0.193	1.72×10^{-158}
RF	Exposure	500	760.409	0.190	2.59×10^{-156}
RF	JPEG	500	762.897	0.191	7.55×10^{-157}
RF	Occlusion	500	596.011	0.149	6.20×10^{-121}

Table 2: Full minus simple hybrid ($\phi_f - \phi_s$) on the fixed participants. Differences and individual 95% bootstrap percentile confidence intervals (CIs) are in percentage points. n_{nz} is the number of nonzero paired participant differences. Raw and globally Holm-adjusted probabilities are from two-sided signed-rank tests.

Model	Context	Difference	Unadjusted 95% CI	n_{nz}	Raw p	Holm p
LR	Clean	+0.050	$[-0.015, 0.121]$	21	0.1335	1.0000
LR	Blur	−0.107	$[-0.247, 0.040]$	12	0.3071	1.0000
LR	Exposure	+0.080	$[-0.040, 0.220]$	9	0.3984	1.0000
LR	JPEG	+0.020	$[-0.120, 0.160]$	13	1.0000	1.0000
LR	Occlusion	+0.014	$[-0.092, 0.111]$	25	0.4805	1.0000
RF	Clean	+0.231	$[0.047, 0.490]$	35	0.0117	0.6216
RF	Blur	+0.223	$[0.033, 0.440]$	15	0.0406	1.0000
RF	Exposure	+0.087	$[-0.073, 0.260]$	11	0.3799	1.0000
RF	JPEG	+0.130	$[-0.020, 0.310]$	7	0.1719	1.0000
RF	Occlusion	+0.094	$[-0.074, 0.261]$	36	0.1324	1.0000

Table 3: Four additional distinct clean TEST full-minus-component-removal Wilcoxon tests (N = 1,000); the angle-removal contrast is already included as full versus simple in Table 2. Differences are in percentage points. Holm adjustment covers all 374 hypotheses.

Classifier	Ablated representation	Difference (pp)	n_{nz}	Holm p
LR	$\phi_{f\setminus A}$	+0.390	71	0.0018
LR	$\phi_{f\setminus \delta}$	+0.919	190	2.53×10^{-11}
RF	$\phi_{f\setminus A}$	+0.467	84	1.47×10^{-5}
RF	$\phi_{f\setminus \delta}$	+0.366	148	0.0227

Table 4: LR, clean: mean participant ranks and upper-triangle global Holm-adjusted Wilcoxon probabilities.

Representation	Mean rank	ϕ_q	ϕ_T	ϕ_A	ϕ_d	ϕ_δ	ϕ_θ	ϕ_s	ϕ_f	ϕ_Z
ϕ_q	6.052	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	1.0000	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_T	5.129	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0001

Representation	Mean rank	ϕ_q	ϕ_T	ϕ_A	ϕ_d	ϕ_δ	ϕ_θ	ϕ_s	ϕ_f	ϕ_Z
ϕ_A	5.649	–	–	–	$< 10^{-4}$	$< 10^{-4}$	0.1751	$< 10^{-4}$	$< 10^{-4}$	0.0026
ϕ_d	4.248	–	–	–	–	1.0000	$< 10^{-4}$	0.0207	0.0054	$< 10^{-4}$
ϕ_δ	4.233	–	–	–	–	–	$< 10^{-4}$	0.0056	0.0038	$< 10^{-4}$
ϕ_θ	6.091	–	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_s	4.092	–	–	–	–	–	–	–	1.0000	$< 10^{-4}$
ϕ_f	4.076	–	–	–	–	–	–	–	–	$< 10^{-4}$
ϕ_Z	5.429	–	–	–	–	–	–	–	–	–

Table 5: LR, blur: mean participant ranks and upper-triangle global Holm-adjusted Wilcoxon probabilities.

Representation	Mean rank	ϕ_q	ϕ_T	ϕ_A	ϕ_d	ϕ_δ	ϕ_θ	ϕ_s	ϕ_f	ϕ_Z
ϕ_q	5.739	–	$< 10^{-4}$	0.0093	$< 10^{-4}$	$< 10^{-4}$	1.0000	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_T	4.975	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0389
ϕ_A	5.562	–	–	–	$< 10^{-4}$	$< 10^{-4}$	0.0541	$< 10^{-4}$	$< 10^{-4}$	0.0560
ϕ_d	4.409	–	–	–	–	1.0000	$< 10^{-4}$	$< 10^{-4}$	0.0008	$< 10^{-4}$
ϕ_δ	4.486	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_θ	6.296	–	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	0.0001
ϕ_s	4.070	–	–	–	–	–	–	–	1.0000	$< 10^{-4}$
ϕ_f	4.115	–	–	–	–	–	–	–	–	$< 10^{-4}$
ϕ_Z	5.348	–	–	–	–	–	–	–	–	–

Table 6: LR, exposure: mean participant ranks and upper-triangle global Holm-adjusted Wilcoxon probabilities.

Representation	Mean rank	ϕ_q	ϕ_T	ϕ_A	ϕ_d	ϕ_δ	ϕ_θ	ϕ_s	ϕ_f	ϕ_Z
ϕ_q	5.610	–	$< 10^{-4}$	1.0000	$< 10^{-4}$	$< 10^{-4}$	1.0000	$< 10^{-4}$	$< 10^{-4}$	0.0004
ϕ_T	4.939	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0295
ϕ_A	5.427	–	–	–	$< 10^{-4}$	$< 10^{-4}$	1.0000	$< 10^{-4}$	$< 10^{-4}$	0.1887
ϕ_d	4.499	–	–	–	–	1.0000	$< 10^{-4}$	1.0000	0.6206	$< 10^{-4}$
ϕ_δ	4.546	–	–	–	–	–	$< 10^{-4}$	0.1654	0.0389	$< 10^{-4}$
ϕ_θ	5.955	–	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	0.0171
ϕ_s	4.410	–	–	–	–	–	–	–	1.0000	$< 10^{-4}$
ϕ_f	4.381	–	–	–	–	–	–	–	–	$< 10^{-4}$
ϕ_Z	5.233	–	–	–	–	–	–	–	–	–

Table 7: LR, jpeg: mean participant ranks and upper-triangle global Holm-adjusted Wilcoxon probabilities.

Representation	Mean rank	ϕ_q	ϕ_T	ϕ_A	ϕ_d	ϕ_δ	ϕ_θ	ϕ_s	ϕ_f	ϕ_Z
ϕ_q	5.657	–	$< 10^{-4}$	0.0317	$< 10^{-4}$	$< 10^{-4}$	1.0000	$< 10^{-4}$	$< 10^{-4}$	0.0001
ϕ_T	4.993	–	–	0.0003	0.0002	0.0001	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.2331
ϕ_A	5.370	–	–	–	$< 10^{-4}$	$< 10^{-4}$	0.7595	$< 10^{-4}$	$< 10^{-4}$	0.1561
ϕ_d	4.607	–	–	–	–	1.0000	$< 10^{-4}$	0.0280	0.0179	$< 10^{-4}$
ϕ_δ	4.611	–	–	–	–	–	$< 10^{-4}$	0.0016	0.0018	$< 10^{-4}$
ϕ_θ	5.786	–	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	0.0123
ϕ_s	4.413	–	–	–	–	–	–	–	1.0000	$< 10^{-4}$
ϕ_f	4.401	–	–	–	–	–	–	–	–	$< 10^{-4}$
ϕ_Z	5.162	–	–	–	–	–	–	–	–	–

Table 8: LR, occlusion: mean participant ranks and upper-triangle global Holm-adjusted Wilcoxon probabilities.

Representation	Mean rank	ϕ_q	ϕ_T	ϕ_A	ϕ_d	ϕ_δ	ϕ_θ	ϕ_s	ϕ_f	ϕ_Z
ϕ_q	5.664	–	$< 10^{-4}$	0.0048	$< 10^{-4}$	$< 10^{-4}$	1.0000	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_T	4.870	–	–	$< 10^{-4}$	0.0030	0.0003	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0179
ϕ_A	5.424	–	–	–	$< 10^{-4}$	$< 10^{-4}$	0.0418	$< 10^{-4}$	$< 10^{-4}$	0.7599
ϕ_d	4.534	–	–	–	–	1.0000	$< 10^{-4}$	0.0093	0.0060	$< 10^{-4}$
ϕ_δ	4.565	–	–	–	–	–	$< 10^{-4}$	0.0022	0.0011	$< 10^{-4}$
ϕ_θ	6.260	–	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	0.0012
ϕ_s	4.256	–	–	–	–	–	–	–	1.0000	$< 10^{-4}$

Representation	Mean rank	ϕ_q	ϕ_T	ϕ_A	ϕ_d	ϕ_δ	ϕ_θ	ϕ_s	ϕ_f	ϕ_Z
ϕ_f	4.248	–	–	–	–	–	–	–	–	$< 10^{-4}$
ϕ_Z	5.179	–	–	–	–	–	–	–	–	–

Table 9: RF, clean: mean participant ranks and upper-triangle global Holm-adjusted Wilcoxon probabilities.

Representation	Mean rank	ϕ_q	ϕ_T	ϕ_A	ϕ_d	ϕ_δ	ϕ_θ	ϕ_s	ϕ_f	ϕ_Z
ϕ_q	7.285	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_T	5.322	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.8332	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_A	4.527	–	–	–	0.1561	1.0000	$< 10^{-4}$	0.0013	$< 10^{-4}$	1.0000
ϕ_d	4.747	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.2881
ϕ_δ	4.484	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	1.0000
ϕ_θ	5.615	–	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_s	4.258	–	–	–	–	–	–	–	0.6216	0.0005
ϕ_f	4.207	–	–	–	–	–	–	–	–	$< 10^{-4}$
ϕ_Z	4.554	–	–	–	–	–	–	–	–	–

Table 10: RF, blur: mean participant ranks and upper-triangle global Holm-adjusted Wilcoxon probabilities.

Representation	Mean rank	ϕ_q	ϕ_T	ϕ_A	ϕ_d	ϕ_δ	ϕ_θ	ϕ_s	ϕ_f	ϕ_Z
ϕ_q	6.577	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_T	5.091	–	–	0.0004	0.0094	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_A	4.730	–	–	–	1.0000	1.0000	$< 10^{-4}$	0.0179	0.0004	0.0730
ϕ_d	4.798	–	–	–	–	0.6546	$< 10^{-4}$	0.0003	$< 10^{-4}$	0.1155
ϕ_δ	4.635	–	–	–	–	–	$< 10^{-4}$	0.0091	0.0002	1.0000
ϕ_θ	5.923	–	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_s	4.437	–	–	–	–	–	–	–	1.0000	1.0000
ϕ_f	4.379	–	–	–	–	–	–	–	–	1.0000
ϕ_Z	4.430	–	–	–	–	–	–	–	–	–

Table 11: RF, exposure: mean participant ranks and upper-triangle global Holm-adjusted Wilcoxon probabilities.

Representation	Mean rank	ϕ_q	ϕ_T	ϕ_A	ϕ_d	ϕ_δ	ϕ_θ	ϕ_s	ϕ_f	ϕ_Z
ϕ_q	6.599	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_T	5.137	–	–	0.0001	0.0012	$< 10^{-4}$	1.0000	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_A	4.780	–	–	–	1.0000	1.0000	$< 10^{-4}$	0.0006	0.0001	1.0000
ϕ_d	4.811	–	–	–	–	0.3699	$< 10^{-4}$	0.0001	$< 10^{-4}$	1.0000
ϕ_δ	4.647	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	0.0003	1.0000
ϕ_θ	5.494	–	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_s	4.469	–	–	–	–	–	–	–	1.0000	0.3300
ϕ_f	4.444	–	–	–	–	–	–	–	–	0.1074
ϕ_Z	4.619	–	–	–	–	–	–	–	–	–

Table 12: RF, jpeg: mean participant ranks and upper-triangle global Holm-adjusted Wilcoxon probabilities.

Representation	Mean rank	ϕ_q	ϕ_T	ϕ_A	ϕ_d	ϕ_δ	ϕ_θ	ϕ_s	ϕ_f	ϕ_Z
ϕ_q	6.530	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_T	5.128	–	–	$< 10^{-4}$	0.0023	$< 10^{-4}$	1.0000	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_A	4.717	–	–	–	1.0000	1.0000	$< 10^{-4}$	0.2622	0.0319	1.0000
ϕ_d	4.836	–	–	–	–	0.6206	$< 10^{-4}$	0.0003	$< 10^{-4}$	0.9756
ϕ_δ	4.712	–	–	–	–	–	$< 10^{-4}$	0.0026	0.0004	1.0000
ϕ_θ	5.365	–	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_s	4.536	–	–	–	–	–	–	–	1.0000	1.0000
ϕ_f	4.511	–	–	–	–	–	–	–	–	0.3536
ϕ_Z	4.665	–	–	–	–	–	–	–	–	–

Table 13: RF, occlusion: mean participant ranks and upper-triangle global Holm-adjusted Wilcoxon probabilities.

Representation	Mean rank	ϕ_q	ϕ_T	ϕ_A	ϕ_d	ϕ_δ	ϕ_θ	ϕ_s	ϕ_f	ϕ_Z
ϕ_q	6.540	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_T	5.080	–	–	$< 10^{-4}$	0.1566	0.0021	0.0339	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_A	4.668	–	–	–	0.9713	1.0000	$< 10^{-4}$	0.1758	0.0317	1.0000
ϕ_d	4.984	–	–	–	–	0.5193	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0091
ϕ_δ	4.841	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.5418
ϕ_θ	5.899	–	–	–	–	–	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ϕ_s	4.330	–	–	–	–	–	–	–	1.0000	1.0000
ϕ_f	4.277	–	–	–	–	–	–	–	–	1.0000
ϕ_Z	4.381	–	–	–	–	–	–	–	–	–

2 Cephalometric Extension

To further characterize the cephalometric extension introduced in the main manuscript, we evaluated the landmark-localization performance of the convolutional detector and the accuracy of anatomical measurements derived from its predicted coordinates. The detector was trained using the provided expert landmark annotations, representing each landmark as a spatial heatmap during training and recovering its two-dimensional coordinates at inference. The evaluation used all 150 radiographs in the official TEST partition, comprising 4,350 anatomical landmarks. Reference coordinates were obtained by averaging the released junior and senior annotation groups, with spatial measurements converted to millimeters using each radiograph's provided scaling factor. The detector returned coordinate predictions for all 29 landmarks in every TEST radiograph. Localization accuracy was assessed using the mean radial error (MRE), defined as the mean Euclidean distance between predicted and reference landmark coordinates. The resulting MRE was 1.462 mm (95% CI: 1.35–1.58 mm). We additionally considered the successful detection rate (SDR), defined as the proportion of landmarks localized within a specified distance of their reference coordinates. Of the 4,350 evaluated landmarks, 79.10% were localized within 2 mm and 95.26% within 4 mm. These results characterize the localization accuracy of the fixed detector under the evaluated protocol.

Beyond individual landmark localization, we investigated whether the extracted anatomical geometry could support interpretable measurements. Specifically, we considered the Sella–Nasion–Point A (SNA) and Sella–Nasion–Point B (SNB) angles, describing the relative positions of the maxilla and mandible, respectively, and their difference (ANB). We additionally evaluated a computational facial-height ratio, defined as

$$100, \frac{d(\mathrm{ANS}, \mathrm{Me})}{d(\mathrm{N}, \mathrm{Me})},$$

where d denotes Euclidean distance, and ANS, Me, and N correspond to the anterior nasal spine, menton, and nasion landmarks.

Comparisons against measurements computed from the reference landmark coordinates yielded mean absolute errors of 1.477°, 1.080°, and 1.059° for SNA, SNB, and ANB, respectively, and 0.901 percentage points for the facial height ratio. These findings illustrate that independently extracted anatomical landmarks can support interpretable geometric measurements. However, neither the clinical validity of these measurements nor the predictive superiority of any particular geometric representation was established by this extension.

References

[1] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937. https://doi.org/10.1080/01621459.1937.10503522.

[2] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. https://doi.org/10.2307/3001968.

[3] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. https://www.jstor.org/stable/4615733.

[4] C. A. Field and A. H. Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):369–390, 2007. https://doi.org/10.1111/j.1467-9868.2007.00593.x.

[5] M. A. Khalid, K. Zulfiqar, U. Bashir, A. Shaheen, R. Iqbal, Z. Rizwan, G. Rizwan, and M. M. Fraz. A benchmark dataset for automatic cephalometric landmark detection and CVM stage classification. *Scientific Data*, 12:1336, 2025. https://doi.org/10.1038/s41597-025-05542-3.